Manipulations avec « FRANTEXT »

Frantext est une base de donnée textuelle, qui regroupe un important corpus de textes français, du XVI^e^ au XX^e^ siècle, saisis sur support informatique. Le corpus est constitué d'environ 3500 œuvres, dont environ 80 % d'œuvres littéraires et 20 % d'ouvrages techniques.

Un logiciel de consultation permet de rechercher, dans une partie du corpus, diverses séquences et leur contexte. Par exemple, si on recherche toutes les occurrences du mot *machination*, on obtiendra une suite d'extraits comme le suivant :

> M332/DUMAS.A PERE / LE COMTE DE MONTE-CRISTO / 1846 page 515 /
> Vous comprenez, mon cher baron, la voie légale est la plus sûre en matière criminelle, c'était peut-être quelque **machination** contre vous.

Ce logiciel est accessible via Internet (aux organismes abonnés), par exemple avec Netscape, et c'est cette interface qui fait l'objet des manipulations d'aujourd'hui.

L'URL correspondant est : `http://zeus.inalf.cnrs.fr/noncateg.htm`,
l'accès à l'interface de consultation (pour les utilisateurs autorisés) se fait en cliquant sur « Activation de Frantext ».

N'hésitez pas à utiliser abondamment les boutons d'aide, le logiciel d'interface étant assez bien (auto-)documenté.

Manipulations

1. *Sélection du corpus.* Menu « Sélection du corpus de travail ». Choisissez auteur (par exemple `Balzac`) ou titre, et cliquez sur « Enregistrer la sélection ». Le nombre de textes sélectionnés est indiqué, et on peut visualiser le corpus (en choisissant « Visualisation de la sélection »).

2. *Recherche dans le corpus.* Menu « Recherche dans les textes ». Selon la taille du corpus choisi, commencer par un mot pas trop fréquent, par exemple `franchement` : remplir la séquence 1, cliquez sur « Cliquez ». Le nombre de résultats est indiqué, visualisez-les en cliquant « Visualisation des résultats ».

3. À partir de la fenêtre de résultats, on peut lancer une autre autre recherche en choisissant « Autre recherche ». Ajouter de la ponctuation pour restreindre le nombre d'occurrences (par exemple en début de phrase, ou en incise...).

4. L'interface dispose d'un outil morphologique, qui permet donc de désigner toutes les formes fléchies d'un lemme. Pour les verbes, il faut utiliser les caractères spéciaux `&c`. Par exemple, recherchez `&cdire` dans le corpus. Comme il y en a trop (environ 15 000 dans tout Balzac), regardez plutôt `&ccoudre`. Repérez les séquences extraites par erreur.

5. Le caractère spécial `&m` s'utilise pour les non-verbes, et permet d'accéder à toutes les formes fléchies (pour les verbes, cela désigne les participes passé et présent). Exemple : `&mvert`.

6. On peut aussi combiner plusieurs mots pour former une séquence, par exemple la séquence `&caimer la vie` permet de trouver toutes les occurrences du verbe *aimer* conjugué, suivi du mot *la* et du mot *vie*.

7. On peut désigner par `&q` un mot quelconque. Chercher par exemple `une &q maison`, ou encore `. &q franchement`.

8. On peut aussi utiliser le caractère `^` avant un mot pour faire référence au complémentaire. Par exemple, `un ^&mbeau jour` désigne la suite de mots *un*, un mot différent de *beau* et de toutes ses formes fléchies, puis *jour*.

9. On peut chercher aussi des co-occurrences en utilisant le cadre « Séquence 2 » (ou 3). Par exemple, on peut chercher toutes les formes de l'expression semi-figée *bouche cousue* en donnant `&mbouche` comme première séquence, `&ccoudre` comme seconde séquence, et en prenant les valeurs par défaut (contexte réduit à la phrase, distance quelconque).

10. Jouez avec les paramètres contexte, distance, ordre entre les séquences pour trouver seulement les occurrences pertinentes des formes de l'expression *prendre une décision*.

11. On peut aussi avoir besoin de manipuler des *listes* de formes. Les listes sont désignées par les caractères spéciaux `&l` suivis du nom de la liste. La liste peut être créée manuellement (par exemple liste des « semi-négations » du français), ou par flexion (par exemple liste des formes du verbe *avoir*), ou par extraction (par *pattern-matching*) du corpus (par exemple liste des mots qui se terminent par *-ement*). Ces listes peuvent être modifiées à la main (cf menus correspondants). À titre d'exercice, ou pourra chercher dans le corpus :
– les phrases négatives
– les formes *avoir* + N (*peur*, *faim*, *envie*...) à l'indicatif ou l'infinitif
– les formes en *-ement* qui sont placées en incise en tête de phrase

12. Avec le menu « Calculs de fréquence », on peut calculer le nombre d'apparitions d'un mot, ou des mots d'une liste (il y a 16 299 occurrences du mot *voyage*), ou extraire des formes et calculer leur fréquence. On peut aussi étudier la distribution des fréquences selon les auteurs ou les œuvres d'un auteur, ou les années. Par exemple, on peut étudier les différentes formes de la négation au cours du temps.

13. Avec le menu « Etude du vocabulaire au voisinage des occurrences d'un mot », que je vous laisse le soin d'explorer, on peut obtenir la liste et les fréquences d'apparition des mots autour d'une forme donnée. Attention à définir un voisinage pas trop important (il peut être défini en phrases ou en mots).

14. Enfin, il est possible de faire des recherches beaucoup plus sophistiquées, en définissant des *grammaires*. Ceci ne peut se faire que dans un corpus préalablement étiqueté (chaque forme est associée avec une catégorie morpho-syntaxique). Pour le moment, seule une partie du corpus Frantext est manipulable sous cette forme. L'URL correspondante est :
`http://zeus.inalf.cnrs.fr/categ.htm`.